DOCUMENT RESUME

ED 175 913 TM 009 553

AUTHOR Coffman, William E.
TITLE Classical Test Development Solutions. Number 23. Iowa Testing Programs Occasional Papers.
INSTITUTION Iowa Testing Programs, Iowa City.
PUB DATE Feb 79
NOTE 36p.; For complete proceedings, see TM 009 232; Paper presented at an Invitational Conference (Reston, Virginia, May 27-30, 1976)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Educationally Disadvantaged; Elementary Secondary Education; Evaluation Needs; *Measurement Goals; Minority Group Children; Norm Referenced Tests; *Program Evaluation; Test Bias; *Test Construction; Testing; Testing Problems; *Test Selection; Test Validity; Theories
IDENTIFIERS Classical Test Theory

ABSTRACT

A classical model for achievement test development was influenced by: the educational tradition of selecting a sample of questions appropriate to a subject and counting the proportion of acceptable responses; the emphasis on obtaining a large and homogeneous sample of questions to insure reliability; and the tradition in psychological testing of scaling performance to meaningful reference groups. The steps in developing tests are: planning; item writing/review; item tryout; final assembly; and norming/scaling. The classical model offers suggestions for selecting tests and evaluating programs for disadvantaged or minority students. Evaluation should extend over two or three years and measure the impact of anticipated achievement gains on self concept and school attitudes. Evidence from the literature or from experts familiar with the local context should indicate that the test is appropriate for program participants, and that it has been screened for bias. To avoid discouraging slow learners or boring high achievers, cut-of-level testing is appropriate, but a careful check on the scaling procedure is crucial. To solve the problem of comparable norms, existing long-term data for a school should be used. Standardized test results should not be ignored if they fail to meet expectations; the data require the more difficult task of interpretations. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED175913

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 23 - February 1979

Classical Test Development Solutions

William E. Coffman

University of Iowa

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at an invitational conference held at Reston, Virginia, May 27-30, 1976 under the sponsorship of the U.S. Office of Education and to be published by CTB/McGraw-Hill in a proceedings volume titled Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation (Michael J. Wargo & Donald Ross Green, Editors)

TM009 553

CLASSICAL TEST DEVELOPMENT SOLUTIONS

William E. Coffman

University of Iowa

When Mike Wargo first asked me if I would prepare a paper for this conference on the topic "Classical Test Development Solutions," my first reaction was to ask, "I wonder what he means by classical test development?" I knew there was a classical test theory with roots going back at least as far as the work of Spearman and formulated in systematic form in Gulliksen's text (1950), but it had been my impression that in most cases actual construction of achievement tests--and perhaps to some extent aptitude and intelligence tests as well--had proceeded from a rational base without much attention to classical theory. I was aware that if there was any tradition guiding the construction of achievement tests it probably rested more on the common sense approach of Rice (1897) and the focus on educational objectives so clearly set forth by Tyler (1934) than on test theory, although there was no doubt that test theory had had some influence on test construction practices. Then, too, there was a stream of tradition growing out of the work of Binet and Simon that was part of the complex picture, and that stream needed to be examined if one were to understand fully the influences on today's standardized achievement tests.

I was reasonably sure that the reason I had been asked to deal with the topic was that I had conducted a number of studies designed to investigate questions of bias in tests. I had always thought of these as involving the application of systematic data collection and analysis in order to clarify the rational categories used by test developers rather than as a substitute for rational categories. On the other hand, I kept

running into statements in the testing literature that suggested at least some people believed there was a well-established traditional, if not classical, test development procedure that had led to all sorts of irrational outcomes--biased tests of intelligence and achievement that labeled little children as inferior on the basis of inappropriate content or on the basis of norms developed with little or no concern for subcultures. These critics seemed to be saying that if we would only substitute criterion referenced tests for norm referenced tests and if the criteria were only chosen to be relevant to the judgment being made, there would be a sudden flooding of the educational world with unbiased sunlight.

I had first become interested in measurement and evaluation while an undergraduate student at Wittenberg University in 1933 and 1934, and I had been close to developments since then, first as a teacher and principal using tests in public schools and later as a specialist in test development, but I thought I'd better take another look at the historical record rather than to depend on my own experience and on my recollections of what competent authorities might have said. Perhaps I had simply been selective in my perceptions. Perhaps there was a classical test development method that contained fatal flaws. On the other hand, perhaps we were experiencing another example of the problem so clearly stated by Farrington: "A great part of the mysticism and superstition of educated men consists of knowledge which has broken loose from its historical moorings." (Levine, 1976, p. 228)

The perception I had been encountering is well expressed in this quotation from a report to the USOE of a project designed to investigate bias in test construction.

> The typical procedure in building standardized achievement and aptitude tests...essentially unchanged over many years (cf. Lord & Novick 1968, Chapter 15; Ruch 1929, Chapter 2)...is first to develop a pool of items meeting various specifications as to form and content. Next these items are given to a sample of individuals.... Various item statistics, such as point biserial correlation (item vs. total score) are calculated and the "best" items are then chosen, with "best" being characterized first and foremost by a high relationship of the item to the total score. Other characteristics such as difficulty and the effectiveness of distracters (in multiple choice tests) are also considered. Most of these latter item characteristics are related to the item-test correlation to some degree. Therefore the items which "discriminate" best, i.e., show the highest relationship to total score, are the ones usually chosen. This in turn means that the characteristics or attributes of the individuals in the tryout sample which are most responsible for differences in total score determine which items tend to be chosen and determine, in effect, what it is the test measures within the range of possibilities available in the item pool. (Green, 1971, p.1)

This is certainly a caricature of the test development process as I had come to know it, and in a later document (Green & Draper, 1972) the

author makes clear that his own conception of the test development process is really much more complex than this. Nevertheless, such critics of standardized tests as Popham (1975) and Walberg (1974) have repeated essentially the same description in all seriousness; so I believe it is incumbent on those of us who have been directly involved in test development for many years to set the record straight. To the extent that achievement tests have been developed following such a procedure, certainly the result has been tests that are reliable but probably "unbalanced" in content and emphasis. On the other hand, has any competent authority advocated such a procedure? I was particularly skeptical of the citations from Ruch and from Lord & Novick and decided to check on them. I found that Lord & Novick, far from supporting the procedure outlined in the quotation, had emphasized the central position of validity in the test development process, called attention to serious limitations of the point biserial, particularly if the test is to be used with groups more or less able than the tryout group, and emphasized the inadequacy of statistical criteria in these words:

> In previous sections we have discussed the problem of item selection in situations where an external criterion is and is not available. The methods presented in those sections involve no item selection algorithm, however, though we mentioned some item indexes that test constructors have found useful in selecting items for a test. It seems fair to say that in much test construction work these indexes receive only moderate attention and the ultimate decision to include or exclude a given test

item depends in large part on the subjective judgment of the test constructor; it is he who decides whether or not the item will contribute to prediction, or whether it measures what he wishes to measure. In the case of achievement tests, particularly, the selection and validation of items is based largely on a judgment of how adequately the set of items spans the domain of interest. The test constructor then uses item statistics primarily to eliminate bad items or to choose the best item from a group of items which he presumes to measure some particular aspect of the domain of interest. (Lord & Novick, 1968, p. 350)

The emphasis on validity is clear also in Ruch's discussion, as is the secondary place of item analysis in deciding which questions to include in a test. He wrote:

Principal methods of validating tests. These may be stated as follows:

1. By judgments of competent persons.
2. By analysis of courses of study or textbooks.
3. By harmonizing with the recommendations of national educational committees or other recognized bodies on curricula, courses of study, minimum essentials, etc.
4. By experimental studies of social utility....
5. By studies of the most frequently recurring errors.
6. By the computation of the percentages of pupils answering each item correctly at each successive age or grade level.

7. By correlation against an outside criterion.
8. By combinations of the above methods. (Ruch, 1929, pp. 28-29)

Notice that Ruch does not even mention calculating the correlation between items and the total score on the pool of items.

A few years later Lindquist was warning against the growing practice of using an internal rather than an external criterion of item selection:

> It is not safe...to select items from an experimental edition for inclusion in a final test solely on the basis of indices of discrimination of the items. In general, the best procedure seems to be to use the index of discrimination solely as a means of identifying seriously defective items, ...but not to employ it in choosing between items showing reasonably high positive indices--the choice in the latter case to be based primarily upon logical rather than statistical considerations. (Hawkes, et al., 1936, p. 54)

The emphasis is mine, although I'm sure Lindquist would approve. Perhaps it would have prevented many instances of malpractice if another quotation from Lindquist had been printed in large type in all the basic textbooks in tests and measurements:

> ...there has been a tendency to give a great deal of prominence to the concept of reliability, not because of its significance, but because it is so easy to measure and describe quantitatively. ...the test with the highest reliability coefficient may be the least valid, while that with the lowest reliability coefficient may be the most valid. (Hawkes, et al., 1936, p. 99)

It seemed quite clear that, at least with respect to achievement tests, the tradition had emphasized the primacy of validity. Was it possible that the misperception had arisen in the context of intelligence testing? To explore this question I turned to the pioneer work of Binet and Simon, originally published in French in 1905, 1908, and 1911, and made available in an English translation in 1916.

Here again I found a clear recognition of the need to base all procedures on careful thinking in relation to purposes. Binet and Simon (1916) differentiated clearly between the "psychological method" and the "pedagogical method" for studying retardation. They wrote:

> Our purpose is to evaluate a level of intelligence. It is understood that we here separate natural intelligence and instruction. It is the intelligence alone that we seem to measure, by disregarding insofar as possible, the degree of instruction which the subject possesses. He should, indeed, be considered by the examiner as a complete ignoramus knowing neither how to read nor write. This necessity forces us to forego a great many exercises having a verbal, literary, or scholastic character. These belong to a pedagogical examination. We believe that we have succeeded in completely disregarding the acquired information of the subject. We give him nothing to read, nothing to write, and submit him to no test in which he might succeed by means of rote learning. In fact we do not even notice his inability to read if the case occurs. (p. 42)

And a little later they wrote:

> It seems to us that in intelligence there is a fundamental faculty, the alteration or the lack of which, is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances.... We shall... bring out in our lists some tests of memory; but so far as we are able, we shall give these tests such a turn as to invite the subject to make absurd replies, and thus under cover of a test of memory, we shall have an appreciation of their judgment. (pp. 42-44)

In other words, central to the work of Binet and Simon was a clear conceptualization of the behavior to be assessed, and a systematic effort to assemble a set of tasks requiring the demonstration of intelligent behavior with a minimum of dependence on instruction.

By the time Monroe (1923) was writing 18 years later, however, this sharp differentiation between the psychological and the pedagogical had been replaced by a clearly pedagogical approach.

> In the measurement of general intelligence it is assumed that all pupils tested have had equal opportunities to acquire the abilities for which the test calls. Although the function of a general intelligence test is to yield measures of native or inherited abilities of pupils, it does this indirectly by measuring certain acquired abilities. Most of the exercises that are included in general intelligence tests make demands

> upon school training and upon the general experiences of pupils. This assumption should be interpreted in terms of the average of all the abilities for which the test calls. It is obvious that, in the case of the abilities called for by a particular sub-test or a particular exercise, all pupils may not have had equal opportunity to acquire them. If the entire test is considered as a whole this assumption is approximately in agreement with actual conditions for children living in the same general environment. (pp. 23-24)

There was, however, the recognition that for a pedagogical test to measure intelligence, the assumption of equal opportunity to learn must be valid. This has proved to be a much easier assumption to state than to demonstrate. Much of the concern about bias in tests of ability in recent years has turned on this issue, and it is not always a simple matter to establish any clear separation of these issues from those one may wish to treat with respect to tests of achievement. In examining solutions to problems of educational program evaluation, we shall need to step gingerly with eyes widely open if we are to avoid this sort of entanglement.

There is another stream flowing from the work of Binet and Simon that needs to be examined if we are to understand clearly the context within which our problem arises. This has to do with the issue of scaling of performance. How does one establish gradients of performance in order to evaluate a program? The solution offered by Simon and Binet was to scale each task in terms of the typical performance of an age group. Then by

arranging tasks in order of difficulty according to this age scale and observing how far along the scale of tasks a pupil could go before encountering tasks he could not complete, one could establish a "mental age" for a pupil.

It is doubtful that Binet and Simon entertained any notion that they were developing a mental age scale in any absolute sense. Certainly their reference groups were children in elementary schools in Paris, and there is no indication that the authors claimed that the samples were necessarily representative of children in general, or of children in Europe, or even of children in France. When a professor in Belgium wrote to complain to Binet that Belgian children in a private school had scored much above age level on the test and that therefore the mental ages appeared to be scaled too low, Binet's response was an acknowledgment that of course such children had had greater opportunity to develop than children in the center of Paris (Binet & Simon, 1916, pp. 316-321). Binet was concerned with the very practical problem of developing a reference scale to help him identify children so far behind their fellows in development that they could not profit from the ordinary instruction, not with establishing for all time a scale of mental development.

It is interesting to note that Binet and Simon did feel that their idea of using units in terms of ages would be meaningful since they noted that teachers "will say without hesitation that such a child is two years or three years retarded," however, they added, "The value of these estimates is as yet undetermined" (Binet & Simon, 1916, pp. 70-71). It well

may be that the lack of success of efforts to eliminate grade equivalent scores or age scores from standardized test batteries may be due to the fact that such units correspond reasonably well to subjective units already in the experience of teachers.

The concept of a set of carefully scaled tasks was picked up by people working on tests of educational achievement, and for a time, at least through the 1920's, one encounters references to educational "tests and scales." Because pupils progressed through the school system of that time by grades rather than strictly by ages it was much easier to obtain representative samples of pupils in a grade rather than pupils of a particular age; and so it became the practice to scale items for achievement tests in terms of grades and tenths of grades rather than in terms of years and months.

Before the onset of large-scale standardized testing, it had been the practice of teachers to construct examinations by assembling sets of questions and scoring papers on the basis of number or percentage of right answers. Thus, it is not surprising that the new scaling procedure was compared to the traditional approach. As evidence accumulated that scores based on the sum of items answered correctly were more reliable than scores based on the scale value of the most difficult item answered correctly, it was these sums of right answers that received the scaling treatment rather than individual items, although items did tend to be roughly scaled through preliminary tryout. By the middle of the 1930's Newkirk and Greene could report:

> Most modern tests are really hybrids resulting from the crossbreeding of these two forms. That is, they are made up of scaled items, but the resulting scores representing accomplishments are expressed in terms of the number of items responded to correctly. (1935, p. 12)

It appears, then, that if there is a classical test development model for achievement testing it has developed out of three streams of experience, the educational tradition of selecting a sample of questions appropriate to a subject of study and counting the number or proportion of acceptable responses, the test theory stream that emphasizes the necessity of obtaining a sufficiently large and homogeneous sample of questions to insure acceptable reliability, and the psychological testing tradition of scaling performance in relation to meaningful reference groups. There was a blurring of the aptitude-achievement distinction as efforts were made (1) to measure general achievement over a variety of skills and content developed in schools and (2) to measure general intelligence on the basis of common learnings, assuming equal opportunity to learn. Performance, at least during the years of rapid growth of intellectual functions, tended to be scaled by comparison with the performance of pupils of typical ages or grades. Central to the model, at least in the minds of the frontier thinkers, was the notion that the validity of an achievement test, in the final analysis rested on the judgment that the tasks included in the test represented stimuli that would elicit performances of educational significance.

Procedures for developing achievement tests

For those not familiar with the procedures, it may be well to review the steps in somewhat greater detail. Logically the steps proceed in a succession over time; in practice, the procedure is much less orderly with information flowing back and forth in complex ways, particularly when the test at the moment is a new form for an established series or a new test in an area where experience has already accumulated. The steps in the process are five: (1) test planning, (2) item writing and review, (3) item tryout, (4) final form assembly, and (5) development of norms and scaling.

Test planning. The planning of a new achievement test is not an undertaking to be approached lightly. Ideally, a completed test plan should include a detailed outline of the content and skills to be measured, illustrations of the types of items to be presented, and a detailed rationale supporting the proposition that if such and such types of exercises are presented the responses of the pupils will show how well they have achieved the intended educational outcomes. The objectives to be included are specified. The sampling frame, in terms both of content and types of skills, is set forth. The relationships between the test plan and actual published statements of curriculum specialists, analyses of widely used textbooks, and other authoritative sources are made explicit. The numbers and difficulty distributions of questions are specified in relation to the time available for testing and the population to be tested. After reading a well-developed test plan, one should be able to follow the thought of the planners step by step. The plan, in other

words, should make sense to an informed critic. Hieronymus and Lindquist (1974) summarize the result of a test plan in these words:

> The Iowa Tests of Basic Skills are criterion-referenced or objective-based in that they were built to reflect progress toward very specific behavioral objectives, stated in terms of what the pupil can do, i.e., in terms of the skills he exhibits. The 217 separate behavioral objectives are listed, item by item, on pages 31-55 of the Teacher's Guide. (p. 9)

Of course, the item by item listing could come only after the test had been assembled, but the basic outline, at least in its draft form, was developed during the planning stage.

Item writing and review. The writing of a good test question is a demanding exercise in communication. The test writer soon comes to accept the fact that "if there is any possible way of misinterpreting the question, somebody will find it." This does not mean that the item writer's task is to remove all ambiguity from the questions. Exercises will appear ambiguous to those who are knowledgeable or who are misinformed (Ebel, 1972, p. 158). On the other hand, it is not an easy task to write a question that all reasonably well-informed people will interpret the same way. I have observed a committee of seven examiners work a full three days formulating to their satisfaction an Advanced Placement Examination consisting of only fifteen essay questions.

The key to effective item writing is to provide for extensive review. The question that appears perfectly straightforward to the original writer may prove not nearly so straightforward to a second authority. In the

professional test development department of a school system, university center, or publisher, the typical item is subjected to a scrutiny difficult to imagine by anyone other than a novelist, or a poet--or another test writer.

In writing and reviewing, the purpose of the test must be kept clearly in mind. If, for example, the test is a reading test, one needs to be sure the item is measuring skill in reading, not memory for what might already have been learned before reading the passages in the test. Can the question be answered without reading the passage? If so, how can it be measuring reading comprehension? On the other hand, can the question be answered simply by perceiving exactly corresponding forms in the reading passage and in the item? If so, is it possibly measuring perceptual skill rather than reading comprehension? Is the item "tricky?" If so, is it measuring skill in avoiding "tricks?" I could go on, but I believe the point is clear; writing items for a test is no task for an amateur. The job requires a clear perception of purposes and a sensitivity to many pitfalls.

Item tryout. The purposes of item tryout are two, to obtain information about the difficulty (scale values) of the questions and to find out whether the communication has been clear, that is, whether the item will function as intended. In the case of the types of tests that are the focus of this conference, tests of basic skills in reading and mathematics, these questions are answered by administering the items to appropriate groups at various stages of development of the skill and studying the patterns of responses. Ideally, the groups should be representative

samples of clearly defined populations that may reasonably be expected to fall at meaningful points along a developmental gradient. In practice, these tend to be samples of pupils that can be arranged in a gradient on some objective basis but that seldom meet ideal specifications, for example, pupils in successive school grades. If the percentages of correct responses to the items are ordered in the same way as the groups, e.g., if items become easier for pupils in higher grades, the items are discriminating, which is simply another way of saying that the response patterns are sensible. Ideally, wrong answer responses ought also to form a hierarchy in the opposite direction, that is, the lower the classification of the group, the greater the percentages of wrong answers. By studying the response patterns for items that do not correspond to expectation, experienced item writers can often achieve insights into what kinds of interpretations may be responsible for the unexpected patterns and what changes in the wording might improve the items. By examining the gradient of percentage of right answers, one can judge the approximate position in the ability scale to place the item. To the extent that the scale is intended to cover a wide developmental range and to assess the ability of subgroups differing widely in background of experience, it is desirable to have tryout information with respect to both the range and the differences in response patterns of the subgroups.

Test theory has developed a variety of indices for summarizing the results of item tryout--the point biserial correlations mentioned earlier, biserial correlations (which are less sensitive to differences in the ability level of tryout groups), parameters of item characteristic curves,

standard score equivalents of percentage right, etc. These are useful shorthand for the initiated, but the indices can never compensate for deficiencies in the tryout groups themselves.

For tests intended for measuring over a wide developmental range it is common practice to administer questions to samples at three adjoining grades and to examine the resulting gradients of response percentages. It is the practice also to divide the samples within grades on the basis of some criterion measure, either the total score on the group of items constituting the pretest or preferably on some appropriate measure independent of the pretest, and to look at response patterns and item statistics. If this examination reveals defects not identified during the item writing and review, the item is discarded or returned to the pool of items awaiting revision; otherwise, it is ready to be considered for the final form of the test.

Test assembly. The assembly of the final form of a standardized test depends on a variety of considerations likely to be overlooked by the inexperienced. Unfortunately, the considerations are likely to be incompatible so that compromises are often necessary. The aim of the test developer is to maximize the utility by achieving an ideal balance among competing criteria. Of primary importance is that the items meet the content characteristics specified in the test plan. The test must be long enough to provide a reliable measure of each pupil's ability yet short enough to be completed in the time that can be spared from the school program. It must maximize the number of functioning items for each pupil yet provide a sufficiently wide spread of difficulty to assess accurately the range of ability within the group to be tested. Ordinarily, for measuring level of achievement one wishes to minimize the

speed factor, but if all pupils are given enough time to answer all questions, the number of questions that can be asked is likely to be too few to achieve the measurement goals for the test. Ultimately, these competing criteria are resolved; but if the test is to serve its purpose, the primary consideration must be that the test provides for an adequate sampling of the outcomes being measured.

For monitoring the growth of skill in reading comprehension and mathematics over the school years, it turns out that a test designed to represent adequately all of the relevant skills and the relevant levels of performance is too long to be administered in the time generally available. Besides, such a test would include items so difficult as to be frustrating for beginners and others so easy as to be insulting to the advanced student. The obvious solution is to divide the test into levels, but it is not an easy matter to decide how to do this. How many levels should there be? How does one decide what level to administer to a particular pupil? How does one insure proper measurement at the points dividing the levels? Different publishers have provided different answers. The ultimate user must decide which solution (all of them are compromises) best meets the requirements of his program.

The test assembler needs to be concerned not only about content but also about format. Any test is artificial in the sense that it presents questions or problems in a setting different from that of ordinary life. The function of format is to minimize the extraneous influences so that the performance will represent as closely as possible what one would observe if the problem were encountered in ordinary life. Directions should

be clear and simple; if pupils cannot be expected to have the skill to read the directions without difficulty, then the directions should be read to the pupils. Within the inherent complexity of the problem or question, the presentation should be as simple and straightforward as possible. Questions of one format should be arranged together to minimize confusion. Print should be legible and pictorial material recognizable. In summary, the test assembly process should supplement the item writing in facilitating clear communication.

Norms and scaling. The final step in the test development process is to administer the test to samples of pupils in order to develop a frame of reference for interpreting the results. For a test of the kind we are discussing, the problem is essentially one of establishing a developmental scale along which one may chart the progress of a pupil or of a group. Since the test typically exists in two or more forms and a number of different levels, the scale cannot be generated directly from the number of questions a pupil gets right on a particular level or form. Ultimately, the development of a scale depends on some mathematical model generated by human thought. The appropriateness of any particular model must be judged by the extent to which it is in agreement with observations of behavior from time to time, from group to group, and from individual to individual.

Experience has shown that when common tests are administered to adjoining school grades, the average score is higher for the higher grade and the spread of scores around the average is greater for the higher grade. Apparently, at least as schools have been operated in the United

States, pupils that have well developed skills at a particular age or grade tend to learn more rapidly than those with less well developed skills. By administering the same level of a test to representative samples of adjoining grades and calculating averages, one can develop reference points for a scale in a manner similar to that used by Binet and Simon. Another way of approaching the problem is to construct a short test covering the full range of ability to be represented by the scale and use the performance of successive grade groups on this common test, and also on the appropriate levels of the test to be scaled, to generate the reference points. (Hieronymus & Lindquist, 1974.) Minor discrepancies in the plots of the reference points for the several distributions are resolved by fitting smooth curves.

It should be noted that a grade-equivalent scale developed in this way will be unique to the particular skill or content area since the relationships from grade to grade differ for the different areas. Thus, one can monitor the progress of a pupil or a class on a grade-equivalent scale, but if one wishes to make comparisons from one area to another at any particular point in time, one will obtain a meaningless answer unless the individual or the group is progressing along the line of averages. To deal with this problem, within grade percentile scores are generated from the cumulative distribution of scores for the sample of the pupils for each grade.

The grade equivalent scale has other limitations, the most serious of which is that grade equivalent scores are easily misinterpreted by the untrained in testing. Various alternate approaches to the development of

scales have been applied, the most common of which is based on the assumption that the several grade samples represent overlapping normal distributions of ability. (Angoff, 1971.)

The samples from which scale scores are derived provide one set of norms that can provide a frame of reference for interpreting the test scores; however, norms, to be most useful, should be based on samples clearly relevant to the interpretation being made. The conscientious test publisher will try to provide a variety of appropriate reference norms; however, the number of different useful norms will undoubtedly always exceed the resources available for generating them. The most useful will usually be local norms developed on the basis of a number of test administrations over a period of years.

Problems and their solutions

It should be obvious from this brief description that the appropriateness of any particular standardized achievement test in evaluating an educational program involving disadvantaged or minority children will depend on there being correspondences between the purposes of the evaluation and the procedures followed in developing the test. As Cronbach (1971) has argued, one does not validate a test but rather an interpretation. A test may be highly valid when used to make one interpretation and extremely invalid when used to make another interpretation. In the short time remaining, let me mention a few problems likely to be faced in evaluating programs involving disadvantaged or minority children and a few of the solutions available through "classical" test development procedures. I'm sure that other speakers will be filling in many of the details.

The problem of the evaluation model. The typical evaluation model that has been applied in evaluating large-scale programs involving disadvantaged and minority students has been an input-treatment-output model. One defines clearly the objectives of the program, obtains a measure of input status with respect to the objectives, applies the treatment, obtains an output measure, and looks for change. In many cases, the time allotted for treatment is short, with input and output measures being obtained within a single school year. The literature on "classical" test development, however, strongly suggests that those learnings that are socially significant and relatively permanent develop slowly under even optimum conditions and are unlikely to be responsive to intensive, short term, treatments. The plant given massive doses of fertilizer may attain luxurious growth only to wither at the first cold breeze. The test we use and the data we collect should be sufficiently extensive over time to check for resistance to cold wind.

Then, too, since there is only a limited amount of time available within a day--or month--or year, concentration on some objectives can be achieved only by placing less emphasis on others. In designing evaluation studies and selecting tests to use, we should heed Robert Pace's (1972) admonition to ask not, "Did we achieve our objectives?" but rather, "What were the consequences?" It isn't sufficient to test for achievement in such basic skills areas as reading and arithmetic. We need to know also what is happening to the pupils' attitudes toward school or to their self concepts. We need to know whether increased growth in reading and arithmetic skills are being achieved at the cost of reduced growth in other aspects of development. We need to be sure that our measures are

sufficiently comprehensive to tell us not only how well we achieved our objectives but also what the overall consequences were.

The problem of domain definition. The typical standardized test of achievement is a machine-scorable, paper and pencil, group administered instrument. In evaluating large-scale educational programs we're likely to be limited to the use of such tests, at least for the comprehensive survey evidence. Basically, however, we're interested in the effects of our educational program on the ongoing, everyday behavior of pupils. Thus, it isn't sufficient to solve our criterion problem by accepting the test performance as an operational definition of what we are trying to accomplish. The "classical" test development process deals with this problem by pooling the judgments of experienced experts, supplemented by a detailed examination of pupil response patterns and special field studies (such as a study of the relationship between scores on an objective test of punctuation, spelling, and usage and ratings of quality of free writing over a period of time). The questions to be asked in choosing tests for evaluating special programs are: Were there, among the experts, any with special knowledge of the population we are studying? Were members of this population represented adequately in the tryout samples or in the field studies? If not, is there sufficient evidence that for the particular skills being measured, such special provisions are unnecessary?

The problem of bias in test form and content. There is an element of truth in the caricature of the traditional test development process with which this paper begins; for a long time the item tryout procedure focused on a single composite sample with no concern for subsamples except as these might be defined by age or grade. The underlying assumption, though

seldom stated, was that differences among pupils were reflecting either native ability or differences in quality of instruction and since the items represented valid objectives in the judgment of experts, differences in scores were telling us what we wanted to know.

But experts develop their expertise by a variety of relevant experiences, and there is always the possibility that some relevant experience may be missing in the pool of experts. The question asked by the dedicated test developer that led to modifications in the test development process over the years might be stated thus: Granted that we want the test to reflect real differences in the abilities measured by the test, is it possible that we have defined the domain in a manner that permits irrelevant factors to fuzz these real differences? Various procedures have been used to investigate this question, all of which involve in one way or another a comparison of the performance of different subgroups of the total populations; and the results have tended to improve our perceptions of the problems involved in developing specifications for tests that minimize the probability of bias (Coffman, 1965). It turns out that boys tend to answer correctly more questions based on reading passages that deal with historical topics while girls do better if the passages deal with literary material. City children typically answer more questions in a given time than rural children. To children in the Southwest, the setting for that first shot at Concord bridge may be much less meaningful than for the child from New England; in contrast, a searching question involving Navaho culture may be more meaningful. If the test is intended to measure deep comprehension of the Revolutionary War experience,

the item on Concord bridge may be legitimate--for both children. If, on the other hand, the test is designed to measure deep comprehension of American culture in general, one had better include both questions--and others reflecting possible special advantages for other subcultures--if one expects to claim the sample of questions is representative. If the test is designed to measure ability to solve typical arithmetic problems and typical "arithmetic problems" are presented in a paper and pencil test, the test may be measuring an irrelevant reading ability for the pupil who cannot read. If such children are included in the population, perhaps the problems should be presented orally.

The "classical" test development procedure provides the necessary tools for providing checks on the judgments of experts; financial and logistic problems often limit their application.

The problem of motivation. Inherent in the "classical" model is the assumption that what one is measuring is what Cronbach (1970) calls maximum performance. It is assumed that each test taker is doing his best, and generally the directions for administering the test are designed to maximize motivation. However, when a single level of a test is administered to a group of widely varying abilities, there is a high probability that less able pupils will become discouraged or that more able ones will become bored.

Some test publishers recommend that this problem be solved by administering an easier or more difficult level of the test to pupils whose developmental level does not correspond to the typical level of a grade or age group. Such a procedure is particularly appropriate for many

pupils in special educational programs, and evidence is accumulating that when pupils can be given tests at the appropriate level of difficulty for them, problems of low motivation, test anxiety, and withdrawal from the competition are minimized (Samuels, 1976).

Unfortunately, the concept that all pupils in a particular grade in school should be learning the same things (and should be tested on them) is so strong that it is difficult to communicate to the lay public, or even to many teachers and administrators, the rationale for "out-of-level" testing. Those responsible for evaluating educational programs for disadvantaged and minority students must take responsibility for solving this communication problem if sound evaluation procedures are to be used. Unfortunately, many find it easier to continue using inappropriate "in-level" tests, which often fail to show the real gains that result from new programs that focus instruction at the appropriate developmental level of the pupils.

The problem of scaling and norms. It requires almost 100 pages of highly technical discussion for Dr. William Angoff (1971) to set forth procedures and problems related to the scaling and norming of tests, and it is not my intention to repeat that discussion here. It is important, however, to be aware of some of Angoff's conclusions when designing studies to assess the effects of special educational programs for disadvantaged and minority students.

The basic question to be answered in these studies is, "How much change has resulted from the application of the special program, and how does this amount compare with what might reasonably be expected had there

been no special program?" One might, of course, design an experiment to answer the question, but sound experimental designs are difficult to apply in the typical educational setting. Often the answer is sought by comparing the performance of the special treatment group with the performance of a norm group. However, Angoff has pointed out that relationships established on one population do not necessarily hold for another population; and populations of disadvantaged and minority students can seldom be considered as equivalent to the population on which a particular test has been scaled and normed.

Ideally, we would like to have our tests scaled and normed on the population we intend to study. Furthermore, we would like to know something about the historical pattern of intellectual development that has characterized this population prior to the introduction of the special program. It may be, for example, that in some schools the programs already in operation have been reasonably successful while in other schools the programs have been less successful. Thus, in comparison with what has been, the special program in one school may appear highly successful while in another school one may observe little improvement, even though both programs may appear equally effective when compared with some overall normative sample.

Angoff (1971, p. 549) also calls attention to problems that may arise as a result of sampling errors in norms. Whereas the sampling errors in individual test scores are random and may be expected to become smaller as data are aggregated, the sampling errors in norms do not become smaller as data are aggregated. Rather, they behave much in the same manner as bias. Thus, if one obtains a norm sample at the beginning of the year and

again at the end of the year by drawing two samples from the same population, whatever sampling error is present in each norm is transmitted directly to every individual comparison and also to any aggregate based on those individual comparisons. Thus, the availability of separate fall and spring norms, unless they are based on the same sample of schools, does not solve the problem of providing us with a reasonable expectation.

One can solve the problem of comparable norms by actually obtaining whatever norms are obtained on a single sample of schools. However, there is evidence that growth patterns within the school year may differ from school to school depending on curricular patterns and grade placement of topics (Beggs and Hieronymus, 1968). Particularly when an input-treatment-output model is used with relatively short treatment periods, one has the uneasy feeling that observed differences, either in favor of or against an experimental educational program, may be reflecting "normal" within school year fluctuations from a norm based on the average over many schools rather than any effect attributable to the special treatment.

Ideally, the evaluation of a special program should involve the use of a long-term data base for a school. Unfortunately, many schools that are eligible for funds to mount special programs do not have such a data base. Where such data are available, a time series analysis (Glass, et al., 1972) is likely to provide the most meaningful insights.

In Summary

From this examination of "classical" test development in its historical perspective, what can one say to the practitioner planning a large-scale evaluation study of programs in reading and mathematics for disadvantaged and minority students? I believe it is safe to say that wholesale rejection of standardized tests or norm referenced interpretations is unwarranted. At the same time, it seems clear that there are pitfalls to avoid and refinements to consider if standardized tests are to provide valid data. At the risk of oversimplifying a very complex issue, a reasonable risk in view of the papers and discussion to follow, I offer the following:

1. Anticipate that only a complex evaluation model will do the job. This means that one must expect each experimental school and each experimental class to be different from every other and worthy of careful study. Assume that the treatment actually applied will be different from the treatment proposed, and make some attempt to determine the extent of the differences.

2. In planning the model, remember that the critical question is "What were the consequences?" not "Did we achieve the objectives in reading and math?" This means that some attempt will be made to collect data on outcomes other than reading and math.

3. Wherever possible, capitalize on existing data. If a school system has been collecting test data for a number of years using a particular battery of tests, a comparison of new data on the same battery,

interpreted in the light of the old data will probably be more meaningful than any interpretation based on a comparison of data from a different battery with normative data for that battery.

4. Be sure that tests you use are appropriate in difficulty for the particular school and the particular level of development of pupils in the program. Usually, this will mean testing "out of level" at least some of the pupils; thus a careful check on the adequacy of the scaling procedure is crucial. It isn't very impressive to report gains by administering a posttest where scores are higher than pretest scores because the posttest has a higher floor and scores were in the chance range on both tests ... or to report no gains when the floor was too high for both pretest and posttest.

5. Be concerned, but not overconcerned, about bias. In choosing tests, ask the publishers for information about how carefully the development procedures included a screening for bias; and when in doubt, set up your own screening procedure. Remember, however, that tests of reading and mathematics are relatively robust in this regard in comparison with tests in some other areas.

6. Validate any tests you use for the particular situation in which they are to be used. This means that judges familiar with the local context should review the tests. It means that test administrations should be observed to determine that pupils are "playing the game." It means that reliability ought to be checked for "populations like this." It means that some attempt will be made to relate performance on the test in that situation to performance in "the real world."

7. Allow enough time for something significant to happen. Significant change in an educational program on a large scale seldom comes in a short time. There are too many variables to be manipulated. Teachers have to be trained in new methods. Lines of communication need to be developed with respect to organizational changes. Those who feel threatened need to be reassured. All these things take time. This does not mean that data useful for formative evaluation should be rejected; rather, it means that any summative evaluation should be based on a minimum of a year and preferably ought to involve two or three years.

8. Finally, remember that "putting it all together" is a job for the human mind, not for a computer. Especially in large-scale evaluation situations, one is seldom faced with data that fulfill the assumptions of experimental designs. This does not mean we should ignore data; rather, it means that we have a more difficult task of interpretation. To explore the implications of this fact is a task I do not propose to pursue at this time.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.), Chapter 15. Washington, DC: American Council on Education, 1971.

Beggs, D. L., & Hieronymus, A. N. Uniformity of growth in the basic skills throughout the school year and during the summer. Journal of Educational Measurement, 1968, 5(2), 91-97.

Binet, A., & Simon, T. (Translated by Elizabeth S. Kite). The development of intelligence in children (The Binet-Simon Scale). Baltimore: Williams & Wilkins Co., 1916.

Coffman, W. E. Principles of developing tests for the culturally different. In Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1966.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.) Educational measurement (2nd ed.) Chapter 14. Washington, DC: American Council on Education, 1971.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder, CO: Laboratory of Experimental Design, University of Colorado, November 1972.

Green, D. R. Racial and ethnic bias in test construction: Final report. Monterey, CA: CTB/McGraw-Hill, 1971.

Green, D. R., & Draper, J. F. Exploratory studies of bias in achievement tests. CTB/McGraw-Hill, September 1972.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (Eds.). The construction and use of achievement examinations. Boston: Houghton Mifflin, 1936.

Hieronymus, A. N., & Lindquist, E. F. Manual for administrators, supervisors, and counselors, Levels edition, Forms 5 & 6, Iowa Tests of Basic Skills. Boston: Houghton Mifflin Co., 1974.

Levine, M. The academic achievement test: its historical context and social functions. American Psychologist, 1976, 31, 228-238.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MS: Addison-Wesley, 1968.

Monroe, W. S. An introduction to the theory of educational measurements. Boston: Houghton Mifflin Co., 1923.

Newkirk, L. V., & Greene, H. A. Tests and measurements in industrial education. New York: Wiley, 1935.

Pace, C. R. Thoughts on evaluation in higher education. Iowa City, IA: The American College Testing Program, 1972.

Popham, J. During open discussion at the AERA Annual Meeting, Washington, DC. March 30-April 3, 1975.

Rice, J. M. The futility of the spelling grind. Forum, 1897, 23, 163-172, 410-419.

Ruch, G. M. The objective or new-type examination: An introduction to educational measurement. Chicago: Scott, Foresman, 1929.

Samuels, S. Individualized testing: a school's point of view. Paper presented at the Twelfth Annual Measurement Conference, Houghton Mifflin Co. & Measurement Research Center, Iowa City, IA, March 30, 1976.

Tyler, R. W. Constructing achievement tests. Columbus, OH: Ohio State University, 1934.

Walberg, H. Educational process evaluation. In Apple, M. W., et al., Educational evaluation: Analysis and responsibility. Berkeley, CA: McCutchan, 1974.